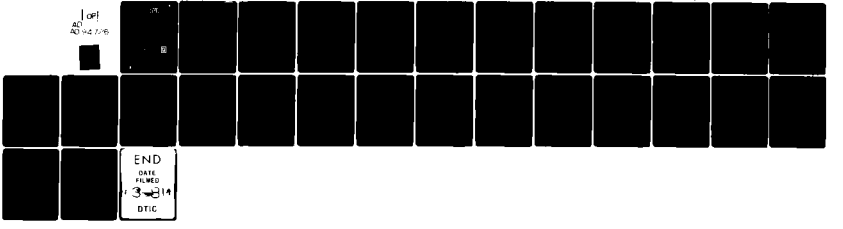
AD-A094 726 MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB F/G 12/1
SOME GENERAL PRINCIPLES FOR THE DUAL PROBLEM TO STATISTICAL CLA--ETC(U)
NOV 80 L K JONES F19628-80-C-0002
UNCLASSIFIED TN-1980-55 ESD-TR-80-229 NL



Technical Note 1980-55

L. K. Jones

Some General Principles for the Dual Problem to Statistical Classification

26 November 1980

Prepared for the Department of the Air Force
under Electronic Systems Division Contract F19628-80-C-0002 by

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology, with the support of the Department of the Air Force under Contract F19628-80-C-0002.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Raymond L. Loiselle

Raymond L. Loiselle, Lt. Col., USAF
Chief, ESD Lincoln Laboratory Project Office

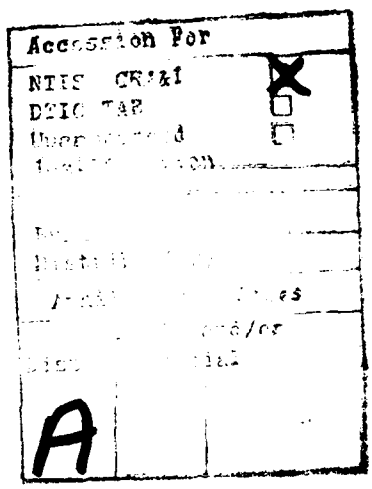
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LINCOLN LABORATORY

SOME GENERAL PRINCIPLES FOR THE DUAL PROBLEM TO STATISTICAL CLASSIFICATION

L. K. JONES

Group 92



TECHNICAL NOTE 1980-55

26 NOVEMBER 1980

LEXINGTON MASSACHUSETTS

ABSTRACT

We consider the design of decision problems which maximize the classification error for a given set of discriminants. A minimax principle is proved, which has applications in discriminant analysis and feature extraction.

SOME GENERAL PRINCIPLES FOR THE DUAL PROBLEM TO STATISTICAL CLASSIFICATION

I. Introduction

In the "two class" problem of statistical classification, we are given two random variables, X_1 and X_2, taking values in R^d. We assume occurrences of type 1 (X_1) and of type 2 (X_2) are mutually exclusive and have prior probabilities of α and $1-\alpha$ respectively. $(0<\alpha<1)$. If x is observed, we need to decide if x is of type 1 or type 2 in such a fashion as to minimize the probability of making an incorrect decision.

If the probability densities (wrt some underlying σ-finite measure ν on R^d) of X_1 and X_2, $p_1(y)$ and $p_2(y)$ were known, we could decide by using the likelihood ratio test.

$$\begin{array}{lll} \dfrac{\alpha p_2(x)}{(1-\alpha)p_1(x)} & > 1 & \text{type 2} \\ & \leq 1 & \text{type 1} \end{array}$$

Unfortunately, these probability densities are often unknown and the problem becomes one of either density estimation or choosing a discriminant function from a class of "feasible" discriminants. There is extensive literature on this subject and we refer the reader to [1], [2], [3].

We define the dual of the two class classification problem as follows: We are given a set Δ of pairs of density functions

for the random variables X_1 and X_2. For which pair $(p_1,p_2)\in\Delta$ is the classification error maximum among all pairs in Δ? For this problem, the structure of Δ is critical. The main difficulties are to transform an applied problem into a mathematical expression for Δ. We consider several examples of dual problems which occur in signal design.

Example 1. The Mixture Problem

Given a density function for X, p(x), and given q_1, q_2,... q_n density functions for Y_1, Y_2,...Y_n find an independent chance device N (Bernoulli r.v.) taking values in {1,2,...n} such that the error for the classification problem X vs. Y_N is maximum. Here $\Delta = \left\{ \left(p(x), \sum_1^n \alpha_i\, q_i(x)\right) : \alpha_i \geq 0; \sum_1^n \alpha_i = 1 \right\}$.

Example 2. The Masking Problem

Let X be a discrete stationary signal of length d. Design a stationary stochastic (independent) signal M of length d such that

(a) the d.c. component = $|EM_i| \leq K_1$

(b) the a.c. power = $VAr\ (M_i) \leq K_2$

and (c) the error is maximized for the problem X+M vs. M.

This problem was discussed in [4] where X and M were restricted to be multivariate normal.

Example 3. A Code Jamming Problem

Let X and Y be discrete real stationary stochastic signals of length d. Find a stationary (independent) signal J such that the error for the problem X+J vs. Y is maximum.

In (4) we shall treat the first example in detail by applying some general principles developed in (2) and (3). We conclude in (5) by considering some theoretical implications to the problem of feature selection.

II. Discriminant Functions and a Minimax Principle

A discriminant function is a map $L:R^d \to R$. Its error is given by

$$\inf_{-\infty < t < +\infty} \left[\alpha \text{ Prob}_1 \{L(X_1) > t\} + (1-\alpha) \text{ Prob}_2 \{L(X_2) \leq t\}\right]$$

for the minimum total error problem. For a given L and a pair (p_1, p_2), denote the above error by $\mathcal{E}_\alpha\left(L; (p_1, p_2)\right)$.

Let $\mathcal{L}$ be a class of discriminant functions. If for a particular dual problem $p_2/p_1 \in \mathcal{L}$ for all $(p_1, p_2) \in \Delta$ (or some other optimal discriminant, i.e., Log p_2 - Log p_1, $p_2 - p_1$, ...), then we are interested in the quantity

$$\max_{(p_1, p_2)} \min_{L} \mathcal{E}_\alpha \left(L; (p_1, p_2)\right)$$

if it exists and $(\bar{p}_1, \bar{p}_2) \in \Delta$ such that

$$\mathcal{E}_\alpha\left(\frac{\bar{p}_2}{\bar{p}_1}\ ;\ (\bar{p}_1,\bar{p}_2)\right) = \min_L\ \mathcal{E}_\alpha\left(L;(\bar{p}_1,\bar{p}_2)\right) =$$

$$\max_{(p_1,p_2)}\ \mathcal{E}_\alpha\left(\frac{p_2}{p_1}\ ;\ (p_1,p_2)\right) = \max_{(p_1,p_2)}\ \min_L\ \mathcal{E}_\alpha\left(L;(p_1,p_2)\right).$$

The first and third equalities follow after a trivial application of Bayes' Theorem. Even if $p_2/p_1 \notin \mathcal{L}$ for all $(p_1,p_2)\in\Delta$ (or no other optimal discriminant), expressions of the form

$$\max_{(p_1,p_2)}\ \min_L\ \mathcal{E}_\alpha\left(L;(p_1,p_2)\right)$$

are appropriate for problems for which only $L \in \mathcal{L}$ are used in the classification problem X_1 vs. X_2. Hence, we continue our discussion for $(p_1,p_2)\in\Delta$ and $L \in \mathcal{L}$. First we show that $\mathcal{E}_\alpha\left(L;(p_1,p_2)\right)$ is concave in (p_1,p_2).

Lemma 1 Assume Δ is convex. Then

$$\mathcal{E}_\alpha\left(L;\left(\gamma p_1+(1-\gamma)\tilde{p}_1,\ \gamma p_2+(1-\gamma)\tilde{p}_2\right)\right) \geq$$

$$\gamma\mathcal{E}_\alpha\left(L;(p_1,p_2)\right) + (1-\gamma)\,\mathcal{E}_\alpha\left(L;(\tilde{p}_1,\tilde{p}_2)\right) \quad \text{for}$$

all $0\leq\gamma\leq 1$ and (p_1,p_2), $(\tilde{p}_1,\tilde{p}_2)\in\Delta$.

Proof. Let $(q_1,q_2) = \gamma(p_1,p_2) + (1-\gamma)(\tilde{p}_1,\tilde{p}_2)$. Then

$$\alpha \text{ Prob}_{q_1} \{L>t\} + (1-\alpha) \text{ Prob}_{q_2} \{L\leq t\}$$

$$= \alpha \left(\gamma \text{ Prob}_{p_1} \{L>t\} + (1-\gamma) \text{ Prob}_{\tilde{p}_1} \{L>t\}\right)$$

$$+ (1-\alpha) \left(\gamma \text{ Prob}_{p_2} \{L\leq t\} + (1-\gamma) \text{ Prob}_{\tilde{p}_2} \{L\leq t\}\right)$$

$$= \gamma \left(\alpha \text{ Prob}_{p_1} \{L>t\} + (1-\alpha) \text{ Prob}_{p_2} \{L\leq t\}\right)$$

$$+ (1-\gamma) \left(\alpha \text{ Prob}_{\tilde{p}_1} \{L>t\} + (1-\alpha) \text{ Prob}_{\tilde{p}_2} \{L\leq t\}\right)$$

$$\geq \gamma \, \mathcal{E}_\alpha \left(L;\ (p_1,p_2)\right) + (1-\gamma) \, \mathcal{E}_\alpha \left(L;\ (\tilde{p}_1,\tilde{p}_2)\right)$$

The result follows by taking the infimum of the above equation over all t.

Note that

$$\mathcal{E}_\alpha \left(\frac{p_2}{p_1},\ (p_1,p_2)\right) = \inf_L \ \mathcal{E}_\alpha \left(L;\ (p_1,p_2)\right)$$

is concave in (p_1,p_2). Hence, $\max_{(p_1,p_2)\in\Delta} \mathcal{E}_\alpha\left(\frac{p_2}{p_1},\ (p_1,p_2)\right)$ could be obtained by methods of convex programming. However, for most problems $\mathcal{E}_\alpha \left(\frac{p_2}{p_1},\ (p_1,p_2)\right)$ is extremely time consuming to evaluate and approximations must be used. We note further that the concavity property fails to hold for the Neyman-Pearson error function at level β, $\mathcal{E}^\beta\left(L;\ (p_1,p_2)\right) = \text{Prob}_2\{L\leq t_\beta\}$ where t_β is

such that $\text{Prob}_1 \{L > t_\beta\} = \beta$. To see this, take $p_1 = p_2 = \tilde{p}_2 = 1$ on $[0,1]$, $\tilde{p}_1 = 1$ on $[-2,-1]$, and $L = x$. Then $\mathcal{E}^{\frac{1}{2}}\left(L; \frac{1}{2}(p_1,p_2) + \frac{1}{2}(\tilde{p}_1,\tilde{p}_2)\right) = 0 < \frac{1}{4} + 0 = \frac{1}{2}\mathcal{E}^{\frac{1}{2}}\left(L; (p_1,p_2)\right) + \frac{1}{2}\mathcal{E}^{\frac{1}{2}}\left(L; (\tilde{p}_1,\tilde{p}_2)\right)$.

Our main result is now a corollary of the following general minimax theorem which seems to be new. The proof, however, is quite standard and will be included for completeness.

Theorem 1. Let C be a subset of a topological space and D a convex compact subset of a topological vector space. Let $f(x,y)$: $C \times D \longrightarrow R$ be a continuous function on $C \times D$, which is concave in y for fixed x. Suppose, further, that there exists a continuous map $x(y)$: $D \longrightarrow C$ s.t. $f(x(y),y) = m(y) = \min_x f(x,y)$. Then

$$\min_x \max_y f(x,y) = \max_y \min_x f(x,y).$$

Proof. For any f on the product of 2 sets C,D we have:

$$f(x,y) \leq \max_y f(x,y) \Rightarrow \min_x f(x,y) \leq \min_x \max_y f(x,y)$$

$$\Rightarrow \max_y \min_x f(x,y) \leq \min_x \max_y f(x,y) .$$

Hence, we need only show

$$\max_y \min_x f(x,y) \geq \min_x \max_y f(x,y)$$

To this end consider y^* such that $m(y^*) = \max_y m(y)$. The existence of y^* is guaranteed since $f(x(y), y)$ is continuous in y. Now

for any t $(0<t\leq 1)$ and $y \in D$,

$$m(y^*) \geq m\left((1-t)y^*+ty\right) = f\left(x\left((1-t)y^*+ty\right),\ (1-t)y^*+ty\right)$$

$$\geq (1-t)\ f\left(x\left((1-t)y^*+ty\right),\ y^*\right) + t\ f\left(x\left((1-t)y^*+ty\right),\ y\right)$$

$$= (1-t)m(y^*) + t\ f\left(x\left((1-t)y^*+ty\right),\ y\right)$$

Therefore, for all $0<t\leq 1$ and any y

$$t\ m(y^*) \geq t\ f\left(x\left((1-t)y^*+ty\right),\ y\right)$$

or $$m(y^*) \geq f\left(x\left((1-t)y^*+ty\right),\ y\right)$$

Letting $t \to 0$ and using the continuity of $f(.,.)$ and $x(\cdot)$ we have

$$m(y^*) \geq f\left(x(y^*),\ y\right) \quad \text{for all } y \in D$$

or $$f\left(x(y^*),\ y^*\right) \geq f\left(x(y^*),\ y\right) \text{ for all } y \in D.$$

Since $f\left(x(y^*),\ y^*\right) \leq f(x,y^*)$ for all $x \in C$, we have for all $x \in C$, $y \in D$

$$f\left(x(y^*),\ y\right) \leq f\left(x(y^*),\ y^*\right) \leq f(x,y^*)$$

Therefore $$\min_x \max_y f(x,y) \leq \max_y f\left(x(y^*),\ y\right) \leq$$

$$f\left(x(y^*),\ y^*\right) \leq \min_x f(x,y^*) \leq \max_y \min_x f(x,y)$$

which was to be shown. The point $(x(y^*),\ y^*)$ is called a <u>Saddle Point</u> of $f(x,y)$.

Corollary 1 Suppose $\Delta = \{(p_1^{\vec{a}}(x), p_2^{\vec{a}}(x)); \vec{a} \in A\}$ is a convex set of pairs of continuous probability densities on R^d such that

$p_i^{\vec{a}}(x) > 0$ for each i, $\vec{a}$, and x

A is a compact subset of R^p for some p

$A \longrightarrow (p_1^{\vec{a}}, p_2^{\vec{a}})$ is continuous in

$$||(f_1, f_2)|| = \sup_{\substack{i=1,2 \\ x \in R^d}} |f_i(x)|$$

Further, let $\mathcal{L}$ be a class of continuous discriminant functions (with the topology of uniform convergence on compact subsets of R^d) with $p_2^{\vec{a}}/p_1^{\vec{a}} \in \mathcal{L}$ (or some other optimal continuous function of $(p_1^{\vec{a}}, p_2^{\vec{a}})$) for each $\vec{a} \in A$. *Then*

$$\min_L \max_{\vec{a}} \mathcal{E}_\alpha\left(L; (p_1^{\vec{a}}, p_2^{\vec{a}})\right) = \max_{\vec{a}} \min_L \mathcal{E}_\alpha\left(L; (p_1^{\vec{a}}, p_2^{\vec{a}})\right)$$

Proof One need only check the continuity of $\mathcal{E}_\alpha(L; (p_1^{\vec{a}}, p_2^{\vec{a}}))$ and the map $A \longrightarrow (p_1^{\vec{a}}, p_2^{\vec{a}}) \longrightarrow \dfrac{p_2^{\vec{a}}}{p_1^{\vec{a}}}$. To verify the former note that $\mathcal{E}_\alpha$ depends (approximately) only on the values of L on a compact set. For the latter note that $A \longrightarrow \dfrac{p_2^{\vec{a}}}{p_1^{\vec{a}}}$ is uniformly continuous in the sup norm restricted to a compact subset of R^d.

Corollary 2 Assume the hypotheses of Corollary 1 with $\mathcal{E}_\alpha$ replaced by $\mathcal{E}^\beta\left(L,(p_1^{\vec{a}},p_2^{\vec{a}})\right) = \text{Prob}_2\ (L \le t_L^{\vec{a}})$ where t_L is such that $\text{Prob}_1\ (L > t_L^{\vec{a}}) = \beta$. Further, assume that $p_1^{\vec{a}} \equiv p_1$ for all $\vec{a} \in A$. (This is the case in Example 1.) Then

$$\min_L \max_{\vec{a}} \mathcal{E}^\beta\left(L;(p_1^{\vec{a}},p_2^{\vec{a}})\right) = \max_{\vec{a}} \min_L \mathcal{E}^\beta\left(L;(p_1^{\vec{a}},p_2^{\vec{a}})\right)$$

Proof Since $p_1^{\vec{a}} \equiv p_1$, $t_L^{\vec{a}}$ depends only on L. Hence, $\mathcal{E}^\beta$ is concave in its second argument and the result follows as in the case of $\mathcal{E}_\alpha$.

We conjecture that the minimax result holds in general for $\mathcal{E}^\beta$.

II. K'th Order Solutions

Suppose for a fixed $L \in \mathcal{L}$ we wish to find $\vec{a}$ s.t. $\mathcal{E}_\alpha\left(L;(p_1^{\vec{a}},p_2^{\vec{a}})\right)$ is maximized. If we assume that the density functions of the $L(p_i^{\vec{a}})$ are completely characterized (differentiably) by their means, variances, ..., K'th moments about the mean and that $\mathcal{E}_\alpha$ is differentiable as a function of the difference of the means of L squared, the variance of L under hypothesis 1, the variance of L under hypothesis 2, ..., the K'th central moment of L under hypothesis 2; then letting $v_i^j(\vec{a})$ = j'th central moment of $L(p_i^{\vec{a}})$ and differentiating $\mathcal{E}_\alpha\left(L;(p_1^{\vec{a}},p_2^{\vec{a}})\right)$ we obtain that a miximal $\vec{a}$ satisfies

$$- \beta^1 \nabla\left[\left(v_2{}^1(\vec{a}) - v_1{}^1(\vec{a})\right)^2\right] + \sum_{i=1}^{2} \sum_{\ell=2}^{K} \beta_i^{\ell} \nabla v_i^{\ell}(\vec{a}) = 0$$

where the $\beta_i{}^j$'s are (usually unknown) scalars corresponding to the partial derivatives of ε_α WRT the various moments. Hence, a maximal $\vec{a}$ is a critical point of the objective function

$$(*) \quad - \beta^1\left(v_2{}^1(\vec{a}) - v_1{}^1(\vec{a})\right)^2 + \sum_{i=1}^{2} \sum_{\ell=2}^{K} \beta_i^{\ell}\, v_i^{\ell}(\vec{a})$$

If ε_α is convex in the above arguments at the critical point in question, then this critical point is a local maximum of (*). Solving for critical points of (*) for various values of $\beta_j{}^j$, allows us to reduce our parameters from p (dimension of A) to 2(K-1). The proofs of the preceding assertions are completely parallel to those in [3] and will hence not be formally presented. We note that the above remains valid if L were allowed to vary with $\vec{a}$. (Provided $L(\vec{a})(p_i{}^{\vec{a}})$ are characterized by their first K central moments.)

As a first example, let $L(\vec{a}) = \ln p_2{}^{\vec{a}}/p_1{}^{\vec{a}}$ and consider the first order solution which is given by the critical points of

$$\left[E_{p_2{}^{\vec{a}}}(\ln p_2{}^{\vec{a}}/p_1{}^{\vec{a}}) - E_{p_1{}^{\vec{a}}}(\ln p_2{}^{\vec{a}}/p_1{}^{\vec{a}})\right]^2 = \left[D(\vec{a})\right]^2$$

where $D(\vec{a})$ is commonly known as the divergence. Since the divergence is always non-negative, the above critical points are the critical points of $D(\vec{a})$. If $\Delta = \{(p_1{}^{\vec{a}}, p_2{}^{\vec{a}})\}$ is convex

then is can be shown (see [5]) that $D(\vec{a})$ is convex in (p_1, p_2) and hence that the critical points are global minima of $D(\vec{a})$. We call such a first order solution a minimal divergence solution.

As a second example, let $L(\vec{a}) = L$ and assume $\beta^1 \neq 0$ for the second order solution (*). Then by dividing (*) by β^1, we see that a second order solution is a critical point of

$$(**) \quad \alpha\, v_1^{\,2}(\vec{a}) + \beta\, v_2^{\,2}(\vec{a}) - \left(v_2^{\,1}(\vec{a}) - v_1^{\,1}(\vec{a})\right)^2$$

where α and β are scalars.

Theorem 2. Suppose Δ is convex and $\alpha \geq 0$, $\beta \geq 0$. Then the critical points of (**) are global maxima of (**).

Proof. We may rewrite (**) as

$$\alpha\, E_{p_1}{}^{\vec{a}}(L^2) + \beta\, E_{p_2}{}^{\vec{a}}(L^2) - \alpha\left(E_{p_1}{}^{\vec{a}}(L)\right)^2$$

$$- \beta\left(E_{p_2}{}^{\vec{a}}(L)\right)^2 - \left(E_{p_2}{}^{\vec{a}}(L) - E_{p_1}{}^{\vec{a}}(L)\right)^2$$

This is clearly concave as a function of $(p_1{}^{\vec{a}}, p_2{}^{\vec{a}})$ and hence any critical point is a global maximum.

Corollary 3 Let Δ be convex and L normal (under each hypothesis) in a neighborhood of a critical point $\vec{a}^1$ of (**) corresponding to a local miximum of $\mathcal{E}_\alpha$. If the error probabilities of each type (at $\vec{a}^1$) are less than ½, then $\vec{a}^1$ is a global maximum of

(**) and $\alpha>0$, $\beta>0$. We call such a solution a <u>maximal variance solution</u> (since we are maximizing a quadratic form in the first two moments with positive coefficients of the variances). Therefore the second order solution for this case is obtained by calculating the error from normal tables for $\vec{a}^1$ which maximizes (*) for a particular choice of α, β and then maximizing over $\alpha>0, \beta>0$.

<u>Proof</u> Since L is normal near $\vec{a}^1$ and the error probabilities of each type are less than .5, the optimal threshold (the minimizing t in the definition of $\mathcal{E}_\alpha$) is between $v_1^1(\vec{a}^1)$ and $v_2^1(\vec{a}^1)$. Clearly

$$\frac{\partial \mathcal{E}_\alpha}{\partial (v_2^1 - v_1^1)^2} < 0 \text{ at } \vec{a}^1$$

and, by the formulae in [3], $\dfrac{\partial \mathcal{E}_\alpha}{\partial v_i^2} > 0$ at $\vec{a}^1$. Hence, $\beta^1>0$, $\beta_1^2>0$, and $\beta_2^2>0$ and a critical point of (*) is a critical point of

$$-\left(v_2^1(\vec{a}) - v_1^1(\vec{a})\right)^2 + (\beta_1^2/\beta^1)\, v_1^2(\vec{a}) + (\beta_2^2/\beta^1)\, v_2^2(\vec{a})$$

which is then a global maximum of

$$\alpha\, v_1^2(\vec{a}) + \beta\, v_2^2(\vec{a}) - \left(v_2^1(\vec{a}) - v_1^1(\vec{a})\right)^2$$

for $\alpha = \beta_1^2/\beta^1 > 0$ and $\beta = \beta_2^2/\beta^1 > 0$.

IV. Application - The Mixture Problem - Stationary Gaussian Case

We now discuss the application of the various techniques of II and III to the problem of Example 1. This problem is of interest for several reasons: for the general dual problem with convex Δ, maximizing error is equivalent to finding the optimal convex combination of the extreme points of Δ and many of the methods of solving Example 1 extend to this more general "mixture" problem; Example 1 occurs often in practical engineering problems -

(i) In a certain communication channel, through which a random signal S may be transmitted, there are several noise signals that can occur. The probabilities of occurrence of each type of signal are small enough that we may assume that no two occur simultaneously. Unfortunately, the relative probabilities of the noise signals are unknown. Solution of this mixture problem yields (by the minimax principle) a detector for S which is optimal in the worst case and performs at least as well in every other case.

(ii) In order to penetrate an enemy radar defense system effectively, the military deploys a variety of decoys as well as a tactical warhead. Assuming the enemy is aware of the statistical radar signatures of the various objects, the military must assign an optimal relative probability to each type of decoy. This is, again, the mixture problem.

A. First Order Solution

Let $p, q_1, q_2, \ldots q_n$ be the stationary, normal densities in R^d of Example 1. We want to determine weights $\alpha_1, \alpha_2, \ldots \alpha_n$; $\alpha_i \geq 0; \sum_1^n \alpha_i = 1$ such that the divergence is minimum. Since the divergence is convex as a function of a pair of probability distributions, it could be minimized by a gradient descent method, provided we are willing to perform many multivariate integrations. If, however, there are mixtures whose density is close to the (normal) density p, we may give approximate expressions for $D(p, \sum_{i=1}^n \alpha_i q_i)$ which are quadratic in α_i and hence easily minimized. More specifically, let p_1 be a stationary normal density with (positive definite) correlation matrix $K = ((k_{ij}))$ and mean $(0,0,\ldots 0)$, let p_2 be a stationary normal density with (positive definite) correlation matrix $K + \Delta$ $\left(\Delta = ((\Delta_{ij}))\right)$ and mean $\vec{m} = (m,m,\ldots m)$, and let Δ_{ij} and m be $0(\varepsilon)$. Then

$$D(p_1,p_2) = \gamma m^2 + \sum_\rho \sum_s q_{\rho s} \Delta_{1\rho} \Delta_{1s} + 0(\varepsilon^3)$$

where $$\gamma = \sum_i \sum_j b_{ij} > 0$$

$$q_{\rho s} = \sum_{j=1}^{d-s+1} \sum_{i=1}^{d-\rho-1} (b_{j\, i+\rho-1}\, b_{i\, j+s-1} + b_{j+s-1\, i+\rho-1}\, b_{ij})$$

$$B = ((b_{ij})) = K^{-1}$$

and $((q_{\rho s}))$ is positive definite.

This is derived in the appendix. Applying the above to our mixture problem with p having mean (0,0,...0) and correlation matrix K and q_i mean $(m_i,m_i,...m_i)$ and correlation matrix Q^i, we need minimize

$$\gamma\left(\sum_{i=1}^{n}\alpha_i m_i\right)^2 + \sum_{\rho}\sum_{s} q_{\rho s}\sum_{i=1}^{n}\alpha_i(Q^i_{1\rho}-k_{1\rho})\sum_{i=1}^{n}\alpha_i(Q^i_{1s}-k_{1s})$$

subject to $\sum_{1}^{n}\alpha_i=1$; $\alpha_i\geq 0$. This is a standard quadratic programming problem.

B. A Second Order Solution

If d is moderately large, then stationarity implies that optimal (quadratic) discriminants will be approximately of the form

$$L = A_0(x_1+x_2+...+x_d) + A_1(x_1^2+x_2^2+...x_d^2) +$$

$$A_2(x_1x_2+x_2x_3+...+x_{d-1}x_d) + A_{d-1}(x_1x_{d-1}+x_{d-1}x_d)$$

$$+ A_d(x_1x_d)$$

In many cases, these discriminants will be approximately normal. Hence, we may use the second order solution for fixed values of $A_0,A_1,...A_d$ and then minimize the associated error over the choice of $A_0,A_1,...A_d$. The above second order solution is equivalent to finding critical points of the one-parameter family of objective functions

$$\beta\ E_2(L^2) - \beta\left[E_2(L)\right]^2 - (E_2L - E_1L)^2$$

where hypothesis 1 has density p and hypothesis 2 density $\sum_{i=1}^{n} \alpha_i q_i$. Note that the first term of (**) is not present in the above objective function since it is constant as a function of α_i. Using the following 2 formulae, valid for x,y,z,w, components of a normal multivariate random variable each with mean m,

$$E(xyz) = m\,E(xy) + m\,E(yz) + m\,E(xz) - 2m^3$$

$$E(xyzw) = E(xy)E(zw) + E(xy)E(yw) + E(xw)E(yz) - 2m^4$$

we calculate

$$E_1(L) = \sum_{\ell=1}^{d} (d-\ell+1)\ A_\ell\ k_{1\ell}$$

$$E_2(L) = \sum_{i=1}^{n} \alpha_i \left[dA_0 m_i + \sum_{\ell=1}^{d} (d-\ell+1)\ A_\ell\ Q^i_{1\ell} \right]$$

$$E_2(L^2) = \sum_{i=1}^{n} \alpha_i \left[A_0{}^2 \sum_{\ell=1}^{d} (n-d+\ell)\ Q^i_{1\ell} + A_0 C^i + D^i \right]$$

where $$C^i = \sum_{a\le b\le c}^{d} m(Q^i_{1b-a+1} + Q^i_{1c-a+1} + Q^i_{1c-b+1} - 2m^2)\ E_{abc}$$

$$D^i = \sum_{a\le b\le c\le g}^{d} (Q^i_{1b-a+1} \cdot Q^i_{1g-c+1} + Q^i_{1c-a+1} \cdot Q^i_{1g-b+1} + Q^i_{1g-a+1} \cdot Q^i_{1c-b+1} - 2m^4)\ F_{abcg}$$

with

$$E_{abc} = \begin{cases} A_{b-a+1} + A_{c-a+1} + A_{c-b+1} & \text{if } a<b<c \\ A_1 + A_{c-a+1} & \text{if } a=b<c \text{ or } a<b=c \\ A_1 & \text{if } a=b=c \end{cases}$$

and

$$F_{abcg} = \begin{cases} A_{b-a+1} \cdot A_{g-c+1} + A_{c-a+1} \cdot A_{g-b+1} + A_{g-a+1} \cdot A_{c-b+1} & \text{if } a<b<c<g \\ A_1 \cdot A_{g-c+1} + A_{g-a+1} \cdot A_{c-a+1} & \text{if } a=b<c<g \\ A_{b-a+1} \cdot A_1 + A_{g-a+1} \cdot A_{g-b+1} & \text{if } a<b<c=g \\ A_1 \cdot A_{g-a+1} + A_{b-a+1} \cdot A_{g-c+1} & \text{if } a<b=c<g \\ A_1 \cdot A_{g-a+1} & \text{if } a<b=c=g \text{ or } a=b=c<g \\ A_1^2 + A_{g-a+1}^2 & \text{if } a=b<c=g \end{cases}$$

Now for $\beta \geq 0$ a critical point of the associated objective function is a global maximum and can hence be determined by standard quadratic programming methods. We note that it might be useful to develop expressions for C^i involving d terms and D^i involving d^2 terms, rendering the computations more feasible for moderately large d.

V. Minimax Feature Extraction

We now describe a purely theoretical application of the minimax principle to feature selection. It is hoped, however, that further research will lead to practical implementation.

Suppose, in the minimum total error classification problem, the (unknown!) densities involved (q_1,q_2) lie in Δ, where Δ is some nontrivial convex set of pairs of possible densities which can be parameterized as in Corollary 1. We may assume that, for a given discriminant function (feature) L, the densities of $L(q_1)$ and $L(q_2)$ can be well enough approximated from sample data that we may consider them known but that the actual higher dimensional densities remain unknown. Then one approach to solving the classification problem is to construct a sequence of discriminant functions $L_0,L_1,\ldots$ whose classification errors decrease. We propose the following sequence: let L_0 be some arbitrary initial discriminant. Let Δ_0 be the set of all pairs of densities $(p_1,p_2)\ \varepsilon\Delta$ such that the density of $L_0(p_i)$ equals that of $L_0(q_i)$. Solve the associated dual problem for Δ_0, obtaining as a solution $(\bar{p}_1,\bar{p}_2)$. Then let $L_1 = \log(\bar{p}_2/\bar{p}_1)$. L_2 is then obtained from L_1 in an identical fashion, etc. We claim that the classification errors for this sequence are non-increasing. To establish this claim, it suffices to prove:

Theorem 3. $$\mathcal{E}_\alpha\left(L_1;(q_1,q_2)\right) \le \mathcal{E}_\alpha\left(L_0;(q_1,q_2)\right).$$

Proof. If the densities of $L_0(\hat{p}_i)$ and $L_0(\tilde{p}_i)$ are identical to that of $L_0(q_i)$, then this density equals the density of $L_0(\gamma\tilde{p}_i+(1-\gamma)\hat{p}_i)$ for any $0\leq\gamma\leq 1$. Hence, Δ_0 is convex and the minimax principle together with Bayes' Theorem imply

$$\mathcal{E}_\alpha\left(L_1;(q_1,q_2)\right) \leq \mathcal{E}_\alpha\left(L_1;(\bar{p}_1,\bar{p}_2)\right) = \mathcal{E}_\alpha\left(\log(\bar{p}_2/\bar{p}_1);(\bar{p}_1,\bar{p}_2)\right)$$

$$\leq \mathcal{E}_\alpha\left(L_0;(\bar{p}_1,\bar{p}_2)\right) = \mathcal{E}_\alpha\left(L_0;(q_1,q_2)\right) .$$

Appendix

Let $L(X) = \log(p_2/p_1)$. Then $L(X) = -\tfrac{1}{2}(X-\vec{m})^t (K+\Delta-m^2)^{-1}(X-\vec{m})$

$$+ \tfrac{1}{2} X^t K^{-1} X + \text{terms not varying in X.}$$

$$(K+\Delta-m^2)^{-1} = K^{-1}\left(I-\Delta K^{-1} + m^2 K^{-1} + (\Delta K^{-1})^2 + 0(\varepsilon^3)\right)$$

Hence, $L(X) = \tfrac{1}{2} X^t K^{-1}\Delta K^{-1} X + \vec{m}^t K^{-1} X - \tfrac{1}{2} X^t K^{-1}(\Delta K^{-1})^2 X$

$$- X^t K^{-1}\Delta K^{-1}\vec{m} - \tfrac{1}{2} X^t m^2 K^{-1} X + 0(\varepsilon^3)$$

$$+ \text{terms not varying in X.}$$

Using the notation $\langle E_2-E_1\rangle Z = E_2 Z - E_1 Z$, we have

$$D(p_1,p_2) = \langle E_2-E_1\rangle L = \langle E_2-E_1\rangle \text{ (L-terms not varying in X).}$$

$$\tfrac{1}{2} X^t K^{-1}\Delta K^{-1} X = \tfrac{1}{2} \sum_j \sum_r \sum_i b_{ri} \sum_\ell \Delta_{i\ell}\, b_{\ell j}\, x_r x_j$$

$$\langle E_2-E_1\rangle \left(\tfrac{1}{2} X^t K^{-1}\Delta K^{-1} X\right) = \tfrac{1}{2} \sum_j \sum_r \left(\sum_i b_{ri} \sum_\ell \Delta_{i\ell}\, b_{\ell j}\right)\Delta_{rj}$$

$$= \tfrac{1}{2} \sum_j \sum_r \left(\sum_\ell \sum_i \Delta_{i\ell}\, b_{ri}\, b_{\ell j}\right)\Delta_{rj}$$

$$= \tfrac{1}{2} \sum_j \sum_r \left(\sum_{\rho=1}^{d} \Delta_{1\rho}\left[\sum_{i=1}^{d-\rho+1} b_{ri}\, b_{i+\rho+j} + \sum_{i=1}^{d-\rho+1} b_{r\, i+\rho-1}\, b_{ij} \right] \right)\Delta_{rj}$$

$$= \tfrac{1}{2} \sum_{\rho=1}^{n} \Delta_{1\rho} \sum_{j} \sum_{r} \sum_{i=1}^{d-\rho+1} \left[b_{ri}\, b_{i+\rho-1\, j} + b_{r\, i+\rho-1}\, b_{ij} \right] \Delta_{ij}$$

$$= \tfrac{1}{2} \sum_{\rho=1}^{n} \Delta_{1\rho} \sum_{i=1}^{d-\rho+1} \sum_{j} \sum_{r} \Delta_{ij} \left(b_{j\, i+\rho-1}\, b_{ir} + b_{r\, i+\rho-1}\, b_{ij} \right)$$

(since B is symmetric)

$$= \sum_{\rho=1}^{n} \Delta_{1\rho} \sum_{i=1}^{d-\rho+1} \sum_{j} \sum_{r} \Delta_{jr}\, b_{j\, i+\rho-1}\, b_{ir}$$

(since Δ is symmetric)

$$= \sum_{\rho=1}^{n} \Delta_{1\rho} \sum_{j} \sum_{r} \Delta_{jr} \sum_{i=1}^{d-\rho+1} b_{j\, i+\rho-1}\, b_{ir}$$

$$= \sum_{\rho=1}^{n} \Delta_{1\rho} \left(\sum_{s=1}^{d} \Delta_{1s} \left[\sum_{j=1}^{d-s+1} \sum_{i=1}^{d-\rho+1} b_{j\, i+\rho-1}\, b_{i\, j+s-1} \right. \right.$$

$$\left. \left. + \sum_{j=1}^{d-s+1} \sum_{i=1}^{d-\rho+1} b_{j+s-1\, i+\rho-1}\, b_{ij} \right] \right)$$

$$= \sum_{\rho=1}^{n} \sum_{s=1}^{n} \Delta_{1\rho}\, \Delta_{1s}\, q_{\rho s}. \quad \text{Further } \langle E_2 - E_1 \rangle \; (\tilde{m}^t K^{-1} X)$$

$$= m^2 \sum_{i} \sum_{j} b_{ij} = \gamma m^2 \quad \text{where } \gamma > 0 \text{ since } K^{-1} \text{ is positive}$$

definite. Finally $\langle E_2-E_1\rangle\ (X^t\ K^{-1}\Delta K^{-1}\vec{m}) = 0(\varepsilon^3)$, $\langle E_2-E_1\rangle$ $(\frac{1}{2}\ X^t K^{-1}(\Delta K^{-1})^2\ X) = 0(\varepsilon^3)$, and $\langle E_2-E_1\rangle(\frac{1}{2}\ X^t m^2 K^{-1} X) = 0(\varepsilon^3)$.

We have now established that $D(p_1,p_2) = \gamma m^2 + \sum_\rho \sum_s q_{\rho s}\ \Delta_{1\rho}\ \Delta_{1s}$ $+ 0(\varepsilon^3)$. It remains to show that $\ll q_{\rho s} \gg$ is positive definite. We present the following analytic proof: Let $a_1, a_2, \ldots, a_d$ be a non-trivial real sequence. Consider Δ defined by $\Delta_{1\rho} = \varepsilon a_\rho$. Since K is positive definite, $K+\Delta$ is positive definite for sufficiently small $\varepsilon > 0$. Setting $\vec{m}=0$, we compute $D(p_1,p_2) > 0$. In fact by passing to a linear space which simultaneously diagonalizes K and Δ, we see that $D(p_1,p_2) = \tau\ \varepsilon^2 + 0(\varepsilon^3)$ for some $\tau > 0$ and ε sufficiently small. It now follows that $\sum_\rho \sum_s q_{\rho s}\ a_\rho\ a_s > 0$.

REFERENCES

[1]. E. Wegman, "Non-Parametric Probability Density Estimation: I. A Summary of Available Methods", Technometrics 14, pgs. 533-546 (1972).

[2]. K. Fukunaga, Introduction to Statistical Pattern Recognition (Academic Press, New York, 1972).

[3]. L. Jones, "K^{th} Order Solutions to the Problem of Finding Optimal Discriminant Functions", submitted to S.I.A.M. J. Appl. Math.

[4]. P. Fishman and L. Jones, "The Minimum Divergence Solution to the Gaussian Masking Problem", submitted IEEE Trans. Inf. Theory.

[5]. P. Fishman and L. Jones, "Convexity Properties of Measures of Class Separation in Statistical Decision Theory", to appear in IEEE Conference on Inf. Theory, February 1981.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

(14) TN-1980-55

(19) **REPORT DOCUMENTATION PAGE**

READ INSTRUCTIONS BEFORE COMPLETING FORM

(18) 1. REPORT NUMBER ESD-TR-80-229	2. GOVT ACCESSION NO. AD A094726
3. RECIPIENT'S CATALOG NUMBER	
(6) 4. TITLE *(and Subtitle)* Some General Principles for the Dual Problem to Statistical Classification •	(1) 5. TYPE OF REPORT & PERIOD COVERED Technical Note
	6. PERFORMING ORG. REPORT NUMBER Technical Note 1980-55
(10) 7. AUTHOR(s) Lee K. Jones	(15) 8. CONTRACT OR GRANT NUMBER(s) F19628-80-C-0002
9. PERFORMING ORGANIZATION NAME AND ADDRESS Lincoln Laboratory, M.I.T. P.O. Box 73 Lexington, MA 02173 (16) 627A	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Program Element No. 63311F Project No. 627A
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Systems Command, USAF Andrews AFB Washington, DC 20331	(11) 12. REPORT DATE 26 November 1980
	13. NUMBER OF PAGES 28
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)* Electronic Systems Division Hanscom AFB Bedford, MA 01731 (12) 28	15. SECURITY CLASS. *(of this report)* Unclassified
	15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

statistical classification

two class problem

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

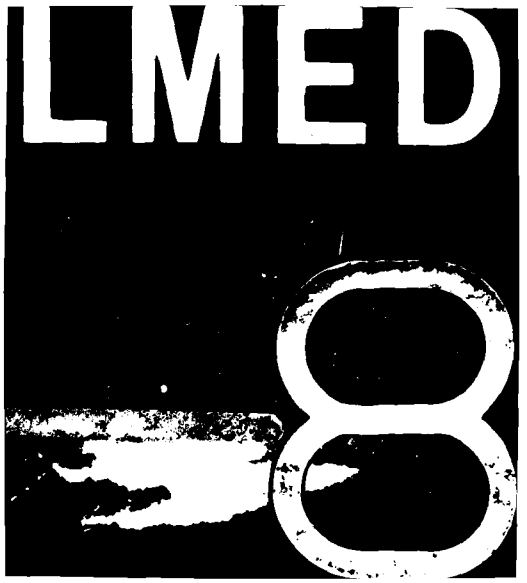
We consider the design of decision problems which maximize the classification error for a given set of discriminants. A minimax principle is proved, which has applications in discriminant analysis and feature extraction.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

207650 YW


LMED
8